

Characterization of open reading frame-expressed sequence tags generated from *Bos indicus* and *B. taurus* mammary gland cDNA libraries

A. F. da Mota^{*,†,‡}, T. S. Sonstegard^{*}, C. P. Van Tassell^{*}, L. L. Shade^{*}, L. K. Matukumalli^{*,§}, D. L. Wood^{*}, A. V. Capuco^{*}, M. A. P. Brito[†], E. E. Connor^{*}, M. L. Martinez[‡] and L. L. Coutinho[‡]

^{*}USDA, ARS Bovine Functional Genomics Laboratory, Beltsville Agricultural Research Institute (BARC) - East, Beltsville, MD, USA. [†]Brazilian Agricultural Research Corporation (EMBRAPA), Gado de Leite/National Dairy Cattle Research Center, Bairro Dom Bosco, Juiz de Fora/MG, Brazil. [‡]Laboratory of Animal Biotechnology, Department of Zootecnia, University of São Paulo/ESALQ, Piracicaba/SP, Brazil.

[§]Bioinformatics and Computational Biology, SCS, George Mason University, Manassas, VA, USA

Summary

Sequence-based gene expression data are used to interpret results from functional genomic and proteomics studies. Although more than 300 000 bovine-expressed sequence tags (ESTs) are available in public databases, a more thorough and directed sampling of the expressed genome is needed to identify new transcripts and improve assembly and annotation of existing transcript sequences. Accordingly, we examined the utility of constructing cDNA libraries synthesized by arbitrarily primed RT-PCR of mRNA from tissues not well represented in the publicly available bovine EST database. A total of 33 cDNA libraries were constructed from healthy and infected mammary gland tissues of Brazilian Gir and Holstein cattle. This series of libraries was used to generate 6481 open reading frame-expressed sequence tags (ORESTES) that assembled into 1798 unique sequence elements of which, 1157 did not significantly match sequence assemblies available in the *Bos taurus* gene index. However, a total of 264 of these 1157 sequence elements aligned with mouse and human expressed sequences demonstrating that ORESTES is an effective resource for discovery of novel expressed sequences in cattle. Furthermore, comparison of the alignment position of bovine ORESTES-derived sequence elements to human gene reference sequences suggested that the priming events for cDNA synthesis more often occurred at the central portion of a transcript, which may have contributed to the relatively high rate of novel sequence discovery.

Keywords bovine transcriptome, expressed sequence tags, mammary gland.

Introduction

Analysis of gene expression from transcriptome- and proteome-based experiments requires sequence data that confers identity of the hybridization target or peptide fragment to a specific gene. Much of the sequence information for human and mouse was generated from expressed sequence tags (EST) with more than 5 million human and 3.6 million mouse EST available for annotation and alignment with

fully assembled genome sequences. The EST collections for these species keep expanding as additional cDNA libraries representing important developmental and disease states are created and sampled to improve existing annotation and ontology of expressed genes.

Greater than 300 000 bovine EST currently reside in the dbEST database of GenBank, with the majority of these EST produced from six normalized cDNA libraries (Smith *et al.* 2001; Sonstegard *et al.* 2002). All cattle-derived transcript sequences were assembled in the *Bos taurus* gene index (BtGI; http://www.tigr.org/tigr-scripts/tgi/T_index.cgi?species=cattle), and human and rodent sequence comparisons were used to guide annotation. The sequence assemblies in BtGI represent gene expression across many of the tissues important to beef and dairy production. However, because most of the EST in these assemblies came

Address for correspondence

T. S. Sonstegard, USDA, ARS Bovine Functional Genomics Laboratory, Bldg 200 Rm 2A Beltsville Agricultural Research Institute (BARC) - East, Beltsville, MD 20705, USA.

E-mail: tads@anri.barc.usda.gov

Accepted for publication 25 March 2004

from cDNA libraries created to maximize sequence discovery by pooling multiple tissues, a number of expressed transcripts specific to critical temporal or physiological states have not been sampled. Certainly, as has been the case for efforts in human beings and mice, additional EST information will be essential to enhance gene identification and assignment of function. Furthermore, EST that correspond to the 3'-ends of genes or that join annotated and non-annotated EST cluster assemblies for the same gene will provide sequence information needed to design oligonucleotides for hybridization arrays.

Consideration should also be given to generating EST information from diverse breeds of cattle. For example, less than 400 accessions in GenBank (as of January 2003) were derived from *B. indicus* cattle breeds that dominate the tropical production areas. Generating EST from these breeds will not only be important for characterizing gene expression patterns of different germplasm sources, but also in broadening the search for expressed polymorphisms related to health and production differences among bovine subspecies.

Consequently, a series of cDNA libraries derived from Holstein and Brazilian Gir mammary tissue were constructed for EST interrogation. As one of the purposes of this study was to maximize sequence discovery in a cost-efficient manner, libraries were generated using a methodology that randomly generates cDNA along the length of a transcript rather than from the 3' extremity of a transcript (Dias Neto *et al.* 2000). These researchers successfully used this methodology to generate nearly 700 000 ORESTES, many of which corresponded to rare transcripts (Camargo *et al.* 2001). In this study, we generated 6481 bovine ORESTES from 33 cDNA libraries. The tissues represented in these libraries included non-lactating mastitic mammary tissues. This shotgun transcript sequencing approach provided an inexpensive method to sample novel bovine sequences with centralized distribution relative to human reference sequences and significant sequence matches to transcripts previously identified only in human beings and rodents.

Materials and methods

Tissue collections

Mammary tissues derived from Brazilian Gir dairy cattle were collected at the Brazilian Agricultural Research Corporation (EMBRAPA) Dairy Cattle Center (Juiz de Fora, Brazil). Four non-pregnant animals undergoing mammary gland involution (30–60 days post-lactation) were selected for post-slaughter tissue collection. Approximately 72 h prior to slaughter, each cow received an intra-mammary infusion in the left rear quarter of 1500 CFU of *Staphylococcus aureus* 284. The bacteria were propagated from a clinical case of mastitis according to protocols used by Schukken *et al.* (1999). Cows were slaughtered at the local abattoir according to animal care and use protocols

approved by EMBRAPA. Tissues from each quarter (1 g per infected and control quarters) were sliced into 5 mm cubes and immediately submersed into RNALater (Ambion, Austin, TX, USA). *Bos taurus* mammary gland was collected from a pre-pubertal Holstein heifer (90 days) at the Beltsville Agricultural Research Center (BARC) abattoir as described by Sonstegard *et al.* (2002).

RNA preparation

Mammary gland mRNA was purified in a three step procedure according to the manufacturer's protocol: (i) extraction using RNeasy mini kits (Qiagen, Valencia, CA, USA); (ii) on-column DNase I digestion of trace contaminating genomic DNA using RNase-Free DNase Set (Qiagen); and (iii) purification of poly-A mRNA with MicroPoly(A) Purist (Ambion). To assess sample integrity after DNase I treatment, total RNA was analysed on an Agilent 2100 bioanalyzer using a RNA 6000 Nano assay (Agilent, Palo Alto, CA, USA). Total RNA samples that displayed no apparent degradation in the corresponding electropherograms (data not shown) and yielded a 28S : 18S ratio greater than 1.9 were selected for mRNA isolation.

Library construction

Bovine cDNA libraries were constructed using a modification of the protocol developed by Dias Neto *et al.* (2000). Briefly, purified mRNA (30 ng) was heat-denatured for 10 min at 65 °C prior to cDNA synthesis with the SuperScript II First-strand Synthesis System (Invitrogen, Carlsbad, CA, USA). The first-strand reactions were carried out at 37 °C for 60 min in a final volume of 10 µl, and 10 pmol of an ORESTES primer was substituted for the poly-dT oligo supplied with the Superscript II kit. cDNA amplification was completed by PCR amplification with the first-strand primer and Ampli-Taq Gold *Taq* DNA polymerase (Applied Biosystems, Foster City, CA, USA) in the presence of one-tenth cDNA yield from the first-strand synthesis reaction as previously described (De Souza *et al.* 2000; Camargo *et al.* 2001). Relative diversity of amplified cDNA products was assessed by electrophoresis through 1% (w/v) agarose gel, and those products between 500 and 1500 bp were excised from gel sample lanes and purified by Qiaquick (Qiagen) for library construction. This size range was chosen to reduce preferential ligation of small cDNA inserts and typically there was no visual evidence of amplified products greater than 1500 bp. The ends of purified cDNA were treated with Klenow fragment (Amersham, Piscataway, NJ, USA) and T4 polynucleotide kinase (Promega, Madison, WI, USA) prior to plasmid ligation and transformation into library competent DH5α or DH10B cells (Invitrogen). Colonies resulting from transformation were picked and arrayed into 384-well plates containing 0.2 ml Luria broth with 100 µg/ml ampicillin for overnight growth.

EST production and analysis

Before glycerol archiving, 2 µl of bacterial culture was removed for PCR amplification according to a protocol made available by Hoffman and colleagues (<http://microarray.cnmcresearch.org/resources.htm>). The only modification was PCR reaction volume was reduced to 10 µl, and 2 µl of each resultant PCR product was treated with 0.01 U of exonuclease I (Epicentre Technologies, Madison, WI, USA) prior to sequencing as described (Sonstegard *et al.* 2002). All sequence trace files were processed, analysed and tracked through EST-PAGE (Matukumalli *et al.* 2003). High-quality ORESTES sequences (>99 consecutive bases of high-quality sequence with Phred >18) after trimming of all vector, single nucleotide repeats ($n > 10$), *Escherichia coli*, and primer sequences were submitted to GenBank dbEST. Bacterial sequences were identified by BLAST to the *E. coli* genome with a threshold *e*-value of 1×10^{-10} . Comparisons were also made to bovine rRNA, mitochondrial DNA, and intron-specific repetitive elements (Sheikh *et al.* 2002) to identify sequences corresponding to potential amplification of hnRNA or contaminating genomic DNA. Unique sequence elements (USEs) from bovine ORESTES were assembled using the default parameters of CAP3 (Huang & Madan 1999).

To evaluate bovine ORESTES relative to transcript diversity and discovery of new transcript sequence, BLAST (Altschul *et al.* 1990) analysis was done against: (1) sequences in BtGI (ver. 8.0) from The Institute of Genomic Research (TIGR; ftp://ftp.tigr.org/pub/data/tgi/bos_taurus/); (2) human and mouse sequences in dbEST; (3) all sequences in nucleotide; and (4) all microbial genomes (2, 3 and 4 are GenBank databases; <http://www.ncbi.nlm.nih.gov>). The threshold for a significant match between bovine sequence elements and GenBank accessions in the different databases was an *e*-value less than 1×10^{-10} . ORESTES-derived sequence elements were placed into a relational database (MYSQL) and aligned by BLAST with the set of full-length human cDNA clones from the Mammalian Gene Collection at the National Institute of Health (<http://mgc.nci.nih.gov/>). A relative percentile position was calculated to normalize alignment position based on the relative distance between the middle base pair of each sequence alignment to the middle nucleotide of its matching reference cDNA.

Results and discussion

We constructed 33 libraries (BARC-EMBRAPA) using cDNA products generated from RT-PCR of poly-A-purified *B. indicus* and *B. taurus* mammary gland mRNA (data not shown). Complete library characteristics are available at <http://estpage.hopto.org>, and primers used for generating the cDNA profiles are listed also in Table 1. The first six libraries were produced from cDNA derived from mRNA of a pre-pubertal Holstein mammary gland and primers

Table 1 Sequences of primers used for cDNA production.

Primer ID	Nucleotide length	Nucleotide sequence (5' → 3')
ESRB2A	23	CCCTGTGACCAAAGACTTGTGTC
ESRB2B	23	GACTCACCTGCTGAATGCTGTGA
ESRA2A	20	TGTGTGGAAGGCATGGTGGA
ESRA1B	21	GGCACCACGTTCTTGCACTTC
ESRR1A	23	GGCATGGCATAACGCTTCTCAGG
ESRR1B	23	AGTACAGCTGTCCGGCCTCCAAC
AM1	16	GGTGCCACTGGACTCG
AM2	16	CTGGACTCGTGCCTCA
AM3	16	CACTGTGCGACTCGTC
AM4	16	CGTCGTACTIONGACGGA
AM5	16	CGCGAGTCGCGCACTT
AM6	16	CAGCACTGTGTCGCGT

(>20 nucleotides) originally designed to amplify oestrogen receptor gene family members. All other libraries were constructed using cDNA derived from Gir cattle and 16-mers of 50–75% guanine and cytosine content to increase the likelihood of generating more diverse cDNA profiles. Varying numbers of cDNA clones from each library were processed for sequencing, and sequence analyses yielded 7817 ORESTES high-quality sequences.

These high-quality sequences were screened to detect *E. coli* and bovine genomic DNA and non-mRNA because of the increased probability of generating potential PCR amplification artefacts and contaminants using low stringency primer annealing conditions for cDNA production. This process removed 1336 sequences (17.1% of total sequences), and most of these were sequences of low complexity (single base repeats) or plasmid sequences with no cDNA insert (Fig. 1). The frequency of sequences derived from bovine genomic DNA (0.4%), rRNA (2.4%), or *E. coli* (5.0%) was relatively low suggesting the removal of non-mRNA was effective. More than half of the *E. coli* sequences were generated from the 321BOV and 322BOV libraries (data not shown), where cDNA production may have occurred in the presence of an exogenous contaminant. The remaining 6481 traces (82.9% of total sequences) were submitted to the dbEST database of GenBank. A total of 5910 submissions were derived from Brazilian Gir cattle, which represents an approximate 20-fold increase in *B. indicus* accessions in GenBank. This sequence information should enhance *in silico* discovery of polymorphisms within transcripts. Once validated, markers developed from these polymorphisms could be used to distinguish *B. indicus* alleles or transcript variants in populations derived from crosses with *B. taurus* breeds.

To further determine the quality of the 6481 ORESTES submissions and the relative efficiency of sequence discovery from the BARC-EMBRAPA libraries, clustering and sequence alignment analyses were carried out. Intra-library assembly generated 730 contigs and 1913 singletons for a total of 2643 USEs. Inter-library assembly produced 1700

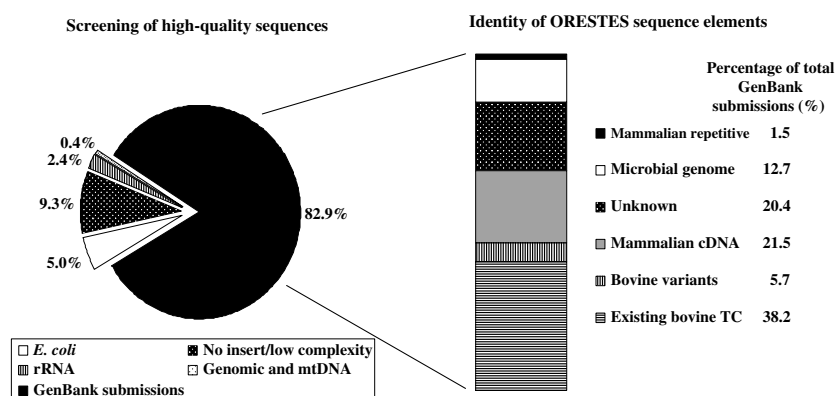


Figure 1 Sequence identities determined for BARC-EMBRAPA ORESTES. Pie graph on the left represents 7817 high-quality sequences and the identities of the slices are shown in the boxed legend. The identities of the 6481 GenBank submissions (largest pie slice) are shown in the bar graph to the right. Sequence identity for ORESTES that did not cluster with existing BtGI TC sequences was determined by BLAST analyses. The ORESTES were grouped into categories as denoted by legend to the right of the bar graph.

USE (556 contigs and 1144 singletons), because 1228 of the intra-library clusters and singletons collapsed into 365 contigs. In this case, the sequence redundancy across libraries was expected and could be attributed to both differential initiation of cDNA synthesis on the same transcript by the various primers and by construction of eight libraries with the same 16-mer (Table 2). Nonetheless, it should be noted that 62% of the contigs formed by clustering across libraries contained EST derived only from a single library. The results of the different BARC-EMBRAPA library assemblies were also compared with the 955 tentative consensus (TC) sequences and 843 singletons created with ORESTES information in BtGI. The total number of ORESTES-containing USE between the BARC-EMBRAPA inter-library assembly and BtGI was nearly identical, and as would be expected, more than 300 singletons from the BARC-EMBRAPA inter-library assembly clustered with TC sequences in BtGI that contained EST from other cDNA libraries. The assembly and annotation information available in BtGI revealed that 2474 bovine ORESTES (38.2% of ORESTES accessions in GenBank) represented additional sequence information for 641 TC sequences co-populated with other bovine EST and cDNA sequences. Another 3124 ORESTES formed 314 new TC sequences that contained only ORESTES sequences, and these sequence elements plus the BARC-EMBRAPA singletons (total of 61.2% of ORESTES accessions) potentially represented previously undiscovered bovine transcripts.

The 1157 sequence elements in BtGI (314 TC sequences and 843 singletons) unique to BARC-EMBRAPA libraries were reassessed for identity using comparative BLAST alignments to three sequence databases, GenBank nucleotide and dbEST for human beings and mice (Fig. 1). Sixty-four of these USE populated by 368 ORESTES (5.7% of accessions), matched existing expressed sequences for cattle. Based on the annotation of the best match and the corresponding *e*-values (most $<1 \times 10^{-40}$) from BLAST comparison (data not shown), most of these 64 USE likely represented allelic variants, splice variants, or paralogs of previously identified bovine genes sequenced from *B. taurus*-derived cDNA. Interestingly, 342 USE comprised 1393

ORESTES (21.5% of accessions), matched sequences from other mammals; and 264 of these matched sequences in GenBank nucleotide and dbEST or matched cDNA sequences in GenBank nucleotide. The other 78 USE aligned only to human genomic DNA sequences, and further investigation is necessary to determine if these alignments represent actual expressed transcripts in cattle. All other USE (34.6% of accessions) fell into one of three categories: (i) unknown (20.4%), (ii) containing mammalian repetitive elements (1.5%), or (iii) being significantly similar to sequences of microbial origin (12.7%). Most ORESTES (90%) in this latter category matched genomic sequences from *Ralstonia solanacearum*, and some of these (5%) also aligned with human or mouse EST at significant but lower *e*-values. All the ORESTES matching *Ralstonia* were derived from libraries constructed with the AM5 primer, which suggests the primer sequence and annealing conditions may have been adequate for amplifying *Ralstonia* contaminants. Moreover, this result underscores both the sensitivity of the ORESTES methodology for generating inserts corresponding to rare mRNA and trace genomic or exogenous DNA contaminants, and the importance of robust screening of ORESTES prior to submission and analysis. In contrast, the production of *Ralstonia*-like sequences may indicate that this microbe species may be a common flora to the mammary gland in Brazilian dairy production systems.

Using the ORESTES assembly statistics, we examined whether substituting 16-mers for cDNA production in place of 23-mers reported in the original protocol by Dias Neto *et al.* (2000) was effective for increasing the diversity of the cDNA products captured into library form. Increasing cDNA clone diversity within a library would lower library production costs and increase throughput by allowing cDNA clones to be processed for sequencing in 384-well plates without generating excessive sequence redundancy. Intra-library assembly revealed that the first six libraries made with different ≥ 20 -mers (300–305BOV) produced approximately one USE for every 2.5 ORESTES after processing an average of 100 cDNA clones per library (Table 2). In contrast, the libraries made with 16-mers (306–356BOV) produced approximately one new sequence for every two

Table 2 Summary of ORESTES sequencing and assembly statistics.

Library ID	Primer	Total EST ¹	Intra-library assembly				Sequence elements unique to library ²
			Contigs	SNG	USE	DI	
300BOV	ESRB2B	70	9	6	15	0.21	15
301BOV	ESRB2A	53	10	15	25	0.47	21
302BOV	ESRA2A	161	28	40	68	0.42	67
303BOV	ESRA1B	86	15	9	24	0.28	18
304BOV	ESRR1B	87	19	13	32	0.37	30
305BOV	ESRR1A	74	15	40	55	0.74	47
306BOV	AM3	14	3	8	11	0.79	8
307BOV	AM2	58	7	45	52	0.90	41
308BOV	AM1	82	13	33	46	0.56	34
309BOV	AM2	184	30	28	58	0.32	49
311BOV	AM3	71	12	17	29	0.41	23
312BOV	AM5	56	7	37	44	0.79	22
313BOV	AM5	47	6	29	35	0.74	12
314BOV	AM5	357	22	62	84	0.24	34
315BOV	AM5	312	50	105	155	0.50	78
316BOV	AM5	347	39	170	209	0.60	114
317BOV	AM5	341	30	94	124	0.36	55
319BOV	Mix	158	17	34	51	0.32	23
321BOV	Mix	144	15	70	85	0.59	52
322BOV	Mix	68	5	29	34	0.50	23
323BOV	Mix	322	20	51	71	0.22	33
324BOV	Mix	224	21	61	82	0.37	29
326BOV	Mix	306	19	39	58	0.19	24
327BOV	Mix	300	21	32	53	0.18	13
328BOV	Mix	301	21	34	55	0.18	18
338BOV	AM5	306	37	151	188	0.61	88
339BOV	AM5	300	35	151	186	0.62	93
351BOV	Mix	316	46	95	141	0.45	87
352BOV	Mix	228	34	113	147	0.64	89
353BOV	Mix	260	38	108	146	0.56	77
354BOV	Mix	261	16	39	55	0.21	23
355BOV	Mix	303	31	73	104	0.34	45
356BOV	Mix	244	39	82	121	0.50	47

Mix, refers to mixing of multiple cDNA derived from different 16-mers; SNG, singleton; USE, unique sequence element; EST, expressed sequence tag; DI, diversity index or USE/EST from within a library.

¹Forty sequence submissions were removed from assembly analysis after further screening for repetitive element content.

²Based on statistics available from BtGI. Totals were only calculated for sequence elements derived from a single library.

ORESTES even after sequencing over 200 cDNA clones per library. This result corroborated with visual observations of cDNA profiles electrophoresed through agarose gels that suggested 16-mers generated more diverse cDNA profiles than ≥ 20 -mers (data not shown). To determine an approximate depth of sequencing from which the sequence discovery rate would significantly decrease below one new sequence for every two ORESTES, eight libraries (312–317BOV, 338BOV and 339BOV) were constructed using the 16-mer, AM5. AM5 consistently yielded the most diverse cDNA profiles of any 16-mer regardless of mRNA source (data not shown). A total of 2066 ORESTES were

generated and assembled from AM5-derived libraries, and the rate of sequence discovery fell below the cumulative average for 16-mer-derived libraries at approximately 750 ORESTES (Fig. 2). Furthermore, the sequence discovery rate was not enhanced by generating some of the AM5-derived libraries with different tissue sources of mRNA. For example, ORESTES from AM5-derived libraries (314BOV and 317BOV) constructed with mRNA from *S. aureus*-infected tissues actually had less cDNA clone diversity than libraries derived from non-infected tissues and contributed very few new sequence elements to inter-library assemblies (Table 2). This lack of sequence diversity may have been

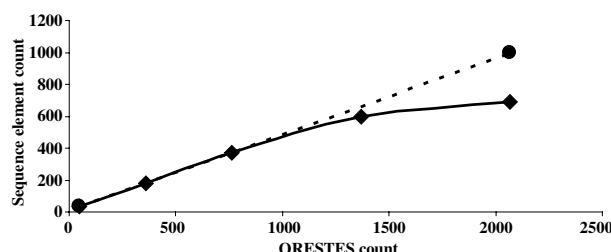


Figure 2 Determination of the rate of sequence discovery for EST from AM5-derived libraries. The number of contigs and singletons (y-axis) relative to the number of ORESTES sequences (x-axis) is plotted as a solid line. The average rate of sequence discovery for all other libraries made with different 16-mers is plotted as a dashed line.

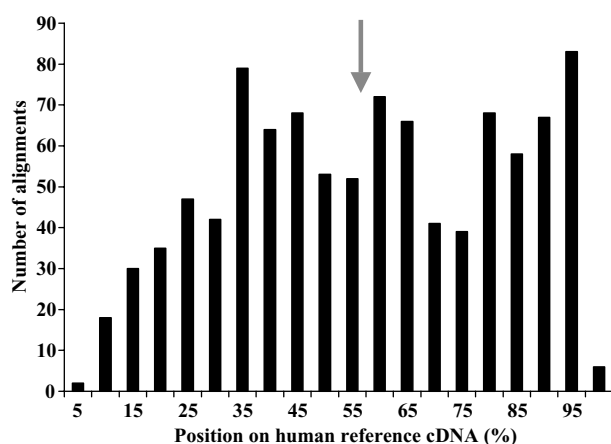


Figure 3 Positional distribution of alignments between unique sequence elements generated from the BARC-EMBRAPA libraries and full-length cDNA sequences of the human reference transcripts. The mean was the 58th percentile and is represented by the grey arrow. The standard deviation was 24 percentile points with a standard error of 0.8 percentile.

caused by the disease state of the tissue collected at slaughter, where loss of synthetic capacity of the epithelial cells was impaired by the extensive damage caused to the gland by infection.

The final analysis was to map the distribution of the bovine ORESTES assemblies relative to full-length reference cDNA, and thereby determine if sequences from the bovine libraries matched the centralized alignment of ORESTES generated from human cDNA libraries (Camargo *et al.* 2001). Because there is a lack of bovine full-length reference sequences, the alignments were carried out against human full-length reference cDNA (Fig. 3). A total of 990 USE (633 contigs and 357 singletons) aligned with the human cDNA, and the mean of the distribution of the central portion of an alignment relative to the centre of the reference cDNA was slightly skewed towards the 3'-end. This result was in general agreement with the results obtained from human ORESTES by Camargo *et al.* (2001).

In conclusion, approximately 27% of the ORESTES-derived sequence elements, not including those sequences with no comparative alignment information, corresponded to bovine transcript information completely novel to BARC-EMBRAPA libraries. Considering the number of existing bovine EST already available for comparison (>300 000) and the amount of total sequence derived from this study, our rate of sequence discovery was relatively robust by generating a novel bovine transcript sequence for every four ORESTES submitted to GenBank. ORESTES methodology appears to be well suited for identifying rare transcripts, targeting the central portion of transcripts, and quick surveying of transcripts from physiologically important tissue samples.

Acknowledgements

We wish to thank Ms Tina Sphon and Dr Marco Machado for superior effort and technical assistance. We also wish to thank Dr John Quackenbush and his laboratory for continued support and management of the TIGR bovine gene index. The authors acknowledge ARS and EMBRAPA for funding of this project through the sponsorship of the Labex-USA program and a CNPq research productivity scholarship received from the National Research Council of Brazil. The authors also acknowledge the support of the EMBRAPA directory, the Brazilian Department of Organization and Development, and the Brazilian Secretariat for International Cooperation. Mention of trade names or commercial products in this article is solely for the purpose of providing specific information and does not imply recommendation or endorsement by the US Department of Agriculture or EMBRAPA.

References

- Altschul S.F., Gish W., Miller W., Myers E.W. & Lipman D.J. (1990) Basic local alignment search tool. *Journal of Molecular Biology* **215**, 403–10.
- Camargo A.A., Samaia H.P., Dias-Neto E. *et al.* (2001) The contribution of 700,000 ORF sequence tags to the definition of the human transcriptome. *Proceedings for the National Academy of Sciences of the United States of America* **98**, 12103–8.
- De Souza S.J., Camargo A.A., Briones M.R. *et al.* (2000) Identification of human chromosome 22 transcribed sequences with ORF expressed sequence tags. *Proceedings for the National Academy of Sciences of United States of America* **97**, 12690–3.
- Dias Neto E., Garcia Correa R., Verjovski-Almeida S., Briones M.R., Nagai M.A., de Silva W., Jr, Zago M.A., Bordin S., Costa F.F. & Goldman G.H. (2000) Shotgun sequencing of the human transcriptome with ORF expressed sequence tags. *Proceedings for the National Academy of Sciences of the United States of America* **97**, 3491–6.
- Huang X. & Madan A. (1999) CAP3: a DNA sequence assembly program. *Genome Research* **9**, 868–77.

- Matukumalli L.K., Grefenstette J.J., Sonstegard T.S. & Van Tassell C.P. (2003) EST-PAGE managing and analyzing EST data. *Bioinformatics* **20**, 286–8.
- Schukken Y.H., Leslie K.E., Barnum D.A., Mallard B.A., Lumsden J.H., Dick P.C., Vessie G.H. & Kehrli M.E. (1999) Experimental *Staphylococcus aureus* intramammary challenge in late lactation dairy cows: quarter and cow effects determining the probability of infection. *Journal of Dairy Science* **82**, 2393–401.
- Sheikh F.G., Mukhopadhyay S.S. & Gupta P. (2002) PstI repeat: a family of short interspersed nucleotide element (SINE)-like sequences in the genomes of cattle, goat, and buffalo. *Genome* **45**, 45–50.
- Smith T.P.L., Grosse W.M., Freking B.A. *et al.* (2001) Sequence evaluation of four pooled-tissue normalized bovine cDNA libraries and construction of a gene index for cattle. *Genome Research* **11**, 626–30.
- Sonstegard T.S., Capuco A.V., White J. *et al.* (2002) Analysis of bovine mammary gland EST and functional annotation of the *Bos taurus* gene index. *Mammalian Genome* **13**, 373–9.